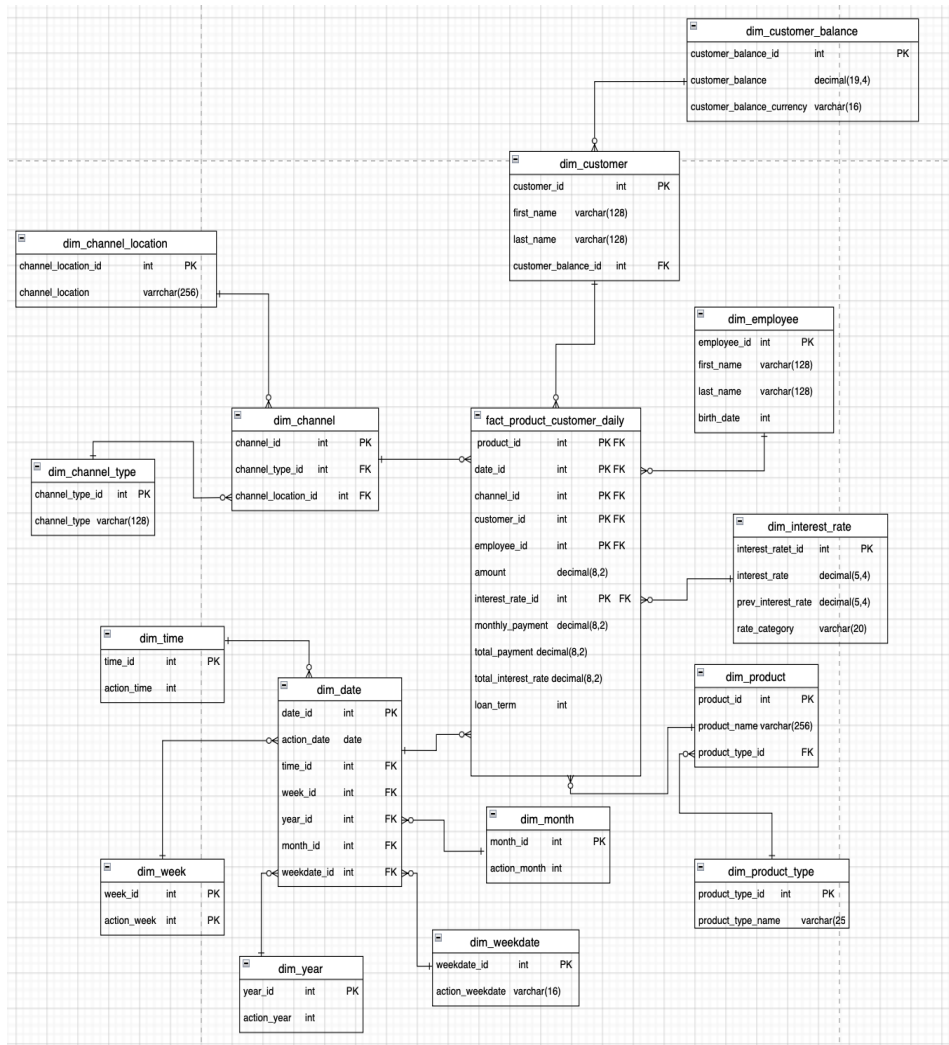


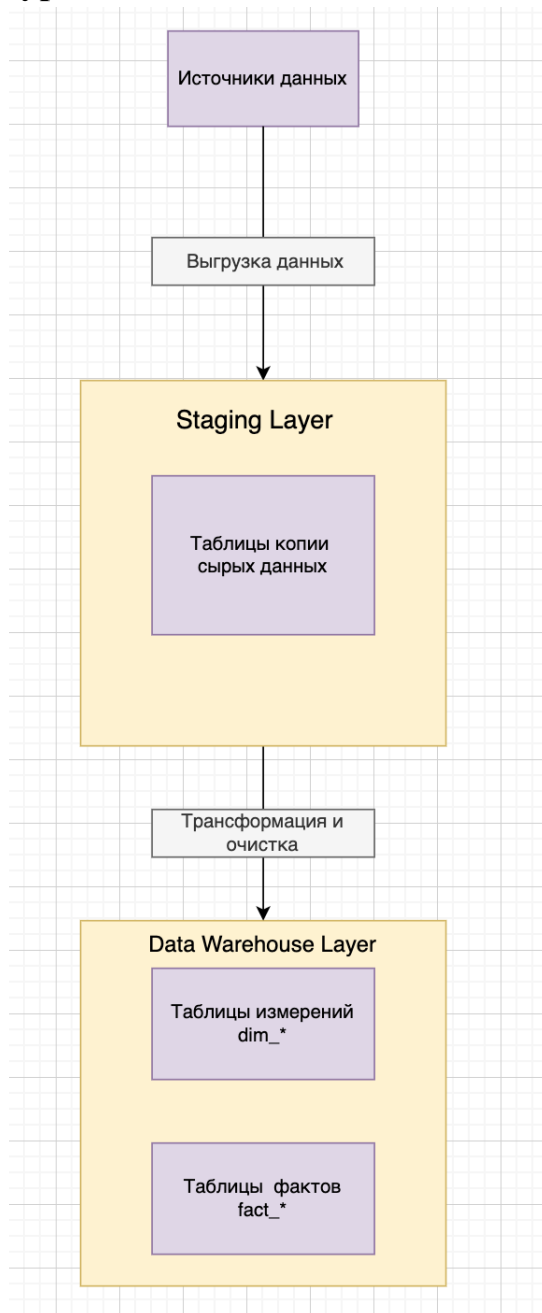
Этап 3. Проектирование потоков данных и ETL-архитектуры

Выполнил:
Воробьев А.И.

ER диаграмма ХД



N1.Общая архитектура потока данных:



Детальное описание этапов потока данных:

1.Источники данных (Data Sources)

данные поступают из различных операционных систем (OLTP):

- CRM-система: Откуда берутся данные о клиентах (customer), сотрудниках (employee) и каналах продаж (channel).

- Финансовая система / ядро банка: Откуда берутся данные о продуктах (product), балансах (customer_balance) и финансовых операциях (факты о продажах, платежах).
- Логи (Logs): Откуда берутся данные о времени и датах действий (хотя измерений времени обычно создаются искусственно в ETL).

2. Промежуточная область (Staging Area)

Это область, куда данные попадают сразу из источников **в сыром виде**, без преобразований.

- **Цель:** Изолировать источники от нагрузки DWH, обеспечить быстрое извлечение и служить точкой восстановления в случае сбоя трансформации.
- **Структура:** Таблицы в Staging повторяют структуру таблиц-источников 1:1. Данные могут добавляться инкрементально или полностью перезаписываться каждый раз.

3. Трансформация и загрузка в DWH (Transformation & Load)

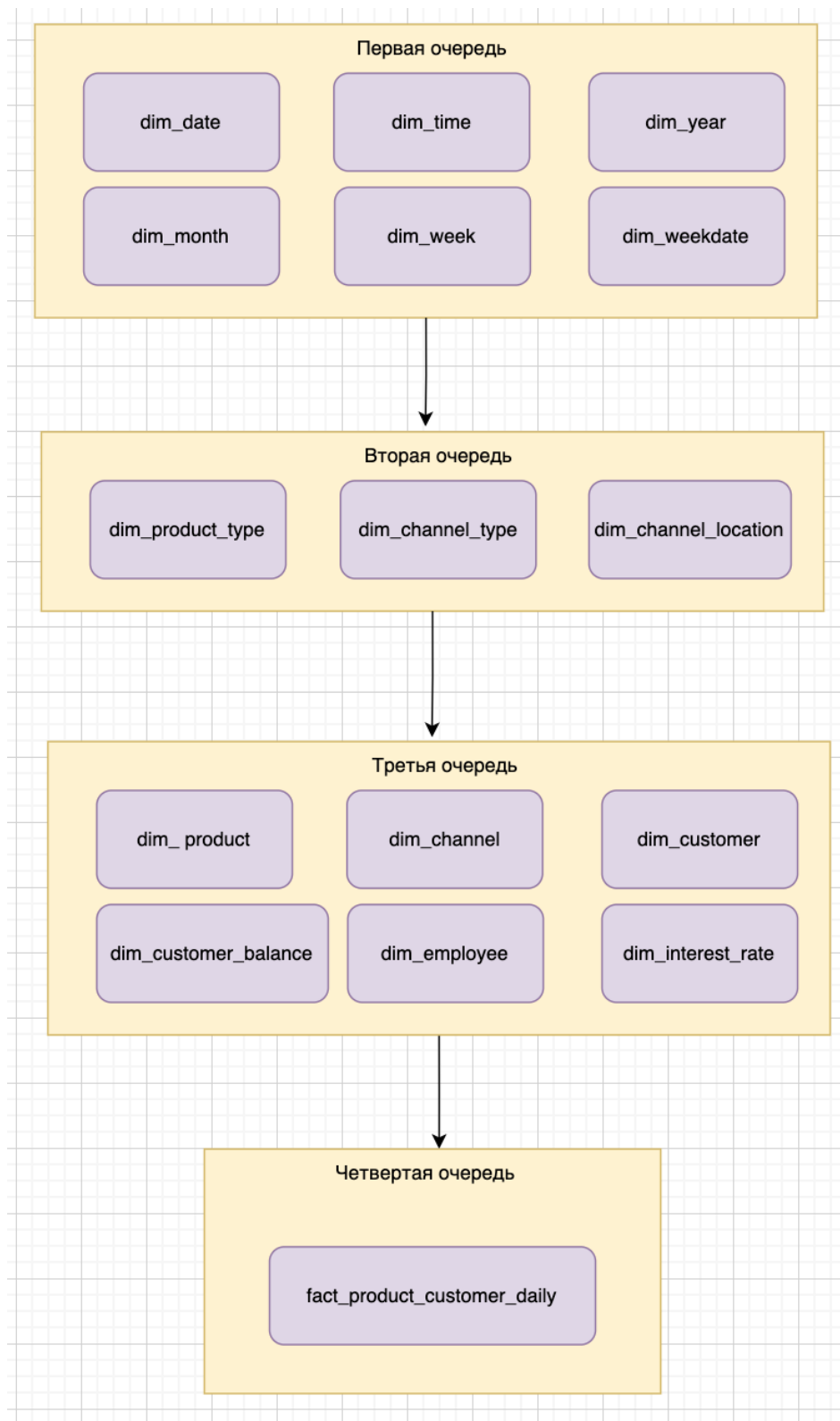
а) Загрузка таблиц измерений (Dimensions):

- **Тип "Снежинка" (SCD) Type 1 (Перезапись истории):** Для атрибутов, где история изменений не важна. Например, dim_channel_location – если изменилось название местоположения, старое можно перезаписать.
- **Тип "Снежинка" (SCD) Type 2 (Сохранить историю):** Для атрибутов, где история изменений критична. Например, dim_customer_balance – необходимо знать, как баланс клиента менялся во времени. Добавляются поля start_date, end_date, is_current.

б) Загрузка таблицы фактов (Facts):

- **Процесс:** Каждая строка из сырых данных о продажах (напр., stg_sales) проходит через серию подзапросов или JOIN'ов для замены "натуральных ключей" (например, имени продукта, даты) на **суррогатные ключи** (ID) из соответствующих таблиц измерений.
- **Даты:** Для замены даты на date_id используется связка таблиц dim_date, чтобы найти корректный ID по полной дате.

N2.Порядок загрузки данных



1. **Календарные измерения** (dim_date, dim_time, dim_year, dim_month, dim_week, dim_weekdate) • Эти таблицы являются базовыми справочниками, которые не зависят от других источников. • Факт и остальные измерения используют их surrogate key (date_id) для привязки записей. Поэтому они загружаются первыми, чтобы при загрузке фактов уже существовали все ключи дат.

2. **Справочные измерения без истории (SCD1):** dim_product_type, dim_channel_type, dim_channel_location • Эти таблицы содержат устойчивые классификаторы (типы продуктов, типы каналов, локации). • Они почти не меняются и не зависят от других сущностей. • Основные измерения (dim_product, dim_channel) используют их как FK. Их загружаем следующими - чтобы основные измерения имели готовые ссылки.

3. **Основные измерения (с историей – SCD2):** dim_product, dim_channel, dim_customer, dim_customer_balance, dim_employee, dim_interest_rate • Эти таблицы описывают основные бизнес-сущности (клиенты, продукты, каналы, сотрудники, ставки). • Они зависят от справочников (например, dim_product → dim_product_type, dim_channel → dim_channel_type, dim_channel_location). • Для них используется стратегия SCD2, поэтому они должны быть загружены до фактов, чтобы факт можно было связать с актуальными surrogate keys. Поэтому на этом шаге формируется полный набор измерений, на который можно опираться при загрузке фактов.

4. **Факт** (fact_product_customer_daily) • Факт - это «центральная таблица», которая хранит количественные показатели (сумма кредита, платежи, срок, ставка). • Он ссылается на все ключевые измерения (клиент, продукт, канал, сотрудник, ставка, дата). • Если измерения ещё не загружены, то факт нельзя корректно загрузить (FK будут отсутствовать). Факт загружается последним - после того как подготовлены все измерения.

N3

Поток данных 1: Измерение dim_date

- **Источник:** Не операционная система. Эта таблица заполняется искусственно.
- **Способ извлечения:** Генерация с помощью скрипта (SQL, Python) или специализированного ETL-инструмента (например, инструмент "Date Dimension" в SSIS или dbt).
- **Преобразования (Transform):**
 - Генерация диапазона дат (например, с 2000-01-01 по 2050-12-31).

- Вычисление атрибутов даты на основе каждой сгенерированной даты:
 - action_date (сама дата)
 - day_of_week (1-7), day_of_month, day_of_year
 - is_weekend (True/False)
 - week_id (номер недели по году)
 - month_id (номер месяца)
 - year_id (год)
 - weekdate_id (ссылка на dim_weekdate, например, 1=Monday)

Целевое поле (Target)	Источник (Source) / Преобразование	Описание
date_id (PK)	GENERATE_SURROGATE_KEY()	Суррогатный ключ, генерируется автоматически.
action_date	date	Сгенерированная дата.
week_id	week_id	Рассчитанный номер недели
year_id	year_id	Рассчитанный год
month_id	month_id	Рассчитанный номер месяца
weekdate_id	weekdate_id	Рассчитанный номер дня недели

Поток данных 2: Измерение dim_product

- **Источник:** Таблица products в OLTP-системе (финансовое ядро банка).
- **Способ извлечения (Extract):**
 - **Первая загрузка:** Полный выгруз (FULL DUMP).
 - **Последующие загрузки: Инкрементальный по дате изменения (INCREMENTAL LOAD).** В таблице-источнике должен быть столбец last_modified_date. Выгружаются только строки, где last_modified_date > даты последней успешной загрузки.
- **Преобразования (Transform):**
 - **Очистка:** Приведение product_name к единому регистру (верхнему). Удаление лишних пробелов.
 - **Обогащение:** Добавление поля product_type_id через JOIN с таблицей product_types из источника или с уже загруженной dim_product_type.
 - **Дедубликация:** Проверка на уникальность по натуральному ключу (например, product_code из источника).
 - **Обработка медленно меняющихся измерений (SCD):** История изменений важна **Тип 2 (Создание новой версии)**

Целевое поле (Target)	Источник (Source) / Преобразование	Описание
product_id (PK)	GENERATE_SURROGATE_KEY()	Суррогатный ключ
product_name	TRIM(UPPER(source.product_name))	Очищенное название продукт
product_type_id (FK)	SELECT pt.product_type_id FROM dim_product_type pt WHERE pt.product_type_name = source.product_type_name	JOIN к целевой размерности для получения суррогатного ключа типа продукта

Поток данных 3: Факты fact_product_customer_daily

- **Источник:** Таблица sales или contracts в OLTP-системе.
- **Способ извлечения (Extract):**

- **Инкрементальная загрузка по дате операции.** Выгружаются все продажи за день, прошедший после последней загрузки (WHERE sale_date = 'YYYY-MM-DD'). Это ежедневная (daily) процедура.
- **Идеальный вариант: CDC (Capture Data Change)** на основе логов транзакций, чтобы захватывать все изменения в режиме, близком к реальному времени.
- **Преобразования (Transform):**
 - **Соединение с измерениями (Lookup):** Это **самое важное преобразование** для фактов. Для каждой строки продажи необходимо найти соответствующие суррогатные ключи во всех связанных измерениях.
 - Найти product_id по коду продукта из источника.
 - Найти customer_id по ID клиента из источника.
 - Найти employee_id по ID сотрудника из источника.
 - Найти channel_id по названию канала из источника.
 - **Найти date_id по дате продажи (sale_date).** Это ключевой шаг.
 - **Валидация:** Отсеивание записей, для которых не найдены ключи в измерениях (например, поступила продажа по неизвестному продукту). Такие записи отправляются в таблицу ошибок для последующего разбора.
 - **Агрегация (опционально):** Если в источнике данные приходят на уровне транзакции, а в целевую таблицу нужно положить ежедневные итоги (что маловероятно, судя по названию таблицы), то проводится агрегация по ключам измерений.

Целевое поле (Target)	Источник (Source) / Преобразование	Описание
product_id (FK)	Lookup('dim_product', source.product_code)	Поиск суррогатного ключа продукта.
product_id (FK)	Lookup('dim_date', source.sale_date)	Поиск суррогатного ключа даты.
channel_id (FK)	Lookup('dim_channel', source.channel_name)	Поиск суррогатного ключа канала.
customer_id (FK)	Lookup('dim_customer_balance	Поиск суррогатного ключа

	', source.customer_id)	клиента.
employee_id (FK)	Lookup('dim_employee', source.employee_id)	Поиск суррогатного ключа сотрудника.
amount	source.amount	Прямое отображение
interest_rate_id	source.interest_rate_id	Прямое отображение.
monthly_payment	source.monthly_payment	Прямое отображение.
total_payment	source.total_payment	Прямое отображение.
total_interest_rate	source.total_interest	Прямое отображение.
loan_term	source.loan_term	Прямое отображение.